Parametric representations of diphthongal formant trajectories of Viennese German /aɛ/

Ewald Enzinger[1]
[1]*Acoustics Research Institute, Austrian Academy of Sciences, Austria*
ewald.enzinger@oeaw.ac.at

Recently, features based on time-dynamic properties of speech segments have been evaluated in the context of forensic speaker comparison. Morrison (2008) used quadratic and cubic polynomial functions fitted to formant trajectories of Australian English /aɪ/ diphthongs. In further studies (Morrison and Kinoshita 2008, Morrison 2009), discrete cosine transform (DCT) components obtained from the formant tracks were added as a second kind of parametric representation. This study uses these methods along with another representation based on B-splines, i.e. pairwise polynomials. They are a generalization of Bézier curves and provide a closer fit to some diphthong formant tracks, at the cost of a higher number of coefficients.

The evaluation is based on diphthongs produced by 30 male Viennese German speakers who were asked to repeat a sentence containing two /aɛ/ segments in the word *kreidebleich*, in different phonetic context and prosodic position. The data is of particular interest because of the monophthongization process in Viennese German, a diachronic process which caused the diphthongs of the Viennese Dialect, /aɛ/ and /aɔ/, to change into the monophthongs /æː/ or /ɛː/ and /ɒː/ or /ɔː/. Speakers of the Viennese Dialect use monophthongs, whereas in Standard Viennese German, monophthongization is a rather gradual process and occurs mostly in prosodically weak positions. The speech samples do not constitute forensically realistic data, as the utterances of each speaker were recorded in only one session, thus neglecting between-session variability.

The likelihood-ratio framework along with the multi-variate kernel density (MVKD) formula developed by Aitken and Lucy (2004) was used for speaker comparison. Sets of coefficient values obtained from the parametric representations of two speakers were used as measurements for suspect and offender samples, while the coefficient values from the remaining speakers were used as reference data.

Evaluations were performed on both /aɛ/ segments separately to avoid effects caused by coarticulation due to disparate phonetic contexts as well as by their different prosodic positions. Although this does not reflect realistic forensic conditions, it enables an assessment of the discriminatory potential of parametric representations based primarily on the diphthongal formant movement.

The results from trials based on coefficient values from two diphthong realizations per speaker showed EER values of 8.3%-12.7% for /aɛ/ in *kreide* and 7%-11% for /aɛ/ in *bleich*, using the first three formant tracks. Tests that included only coefficients derived from F2 and F3 returned values of 9.7%-15.7% and 8.1%-13.3%, respectively.

References

Aitken, C. G. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics,* **53**, 109-122.

Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *Speech, Language, and the Law,* **15**, 249-266.

Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America,* **125**, 2387-2397.

Morrison, G. S. and Kinoshita, Y. (2008). Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. *Proceedings of Interspeech 2008 incorporating SST'08*, 1501-1504.